Appeared in the 2002 International Conference on Automatic Face and Gesture Recognition

The CMU Pose, Illumination, and Expression (PIE) Database

Terence Sim, Simon Baker, and Maan Bsat

The Robotics Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213

Abstract

Between October 2000 and December 2000 we collected a database of over 40,000 facial images of 68 people. Using the CMU 3D Room we imaged each person across 13 different poses, under 43 different illumination conditions, and with 4 different expressions. We call this database the CMU Pose, Illumination, and Expression (PIE) database. In this paper we describe the imaging hardware, the collection procedure, the organization of the database, several potential uses of the database, and how to obtain the database.

1 Introduction

People look very different depending on a number of factors. Perhaps the three most significant factors are: (1) the pose; i.e. the angle at which you look at them, (2) the illumination conditions at the time, and (3) their facial expression; i.e. whether or not they are smiling, etc. Although several other face databases exist with a large number of subjects [Philips *et al.*, 1997], and with significant pose and illumination variation [Georghiades *et al.*, 2000], we felt that there was still a need for a database consisting of a fairly large number of subjects, each imaged a large number of times, from several different poses, under significant illumination variation, and with a variety of facial expressions.

Between October 2000 and December 2000 we collected such a database consisting of over 40,000 images of 68 subjects. (The total size of the database is about 40GB.) We call this database the CMU Pose, Illumination, and Expression (PIE) database. To obtain a wide variation across pose, we used 13 cameras in the CMU 3D Room [Kanade *et al.*, 1998]. To obtain significant illumination variation we augmented the 3D Room with a "flash system" similar to the one constructed by Athinodoros Georghiades, Peter Belhumeur, and David Kriegman at Yale University [Georghiades *et al.*, 2000]. We built a similar system with 21 flashes. Since we captured images with, and without, background lighting, we obtained $21 \times 2 + 1 = 43$ different illumination conditions. Finally, we asked the subjects to pose with several different "expressions." In particular, we asked them to give a neutral expression, to smile, to blink (i.e. shut their eyes), and to talk. These are probably the four most frequently occurring "expressions" in everyday life.

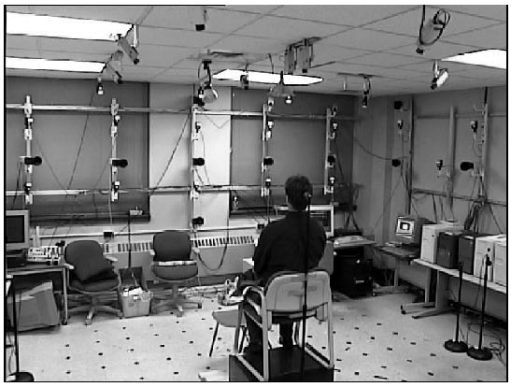


Figure 1: The setup in the CMU 3D Room [Kanade *et al.*, 1998]. The subject sits in a chair with his head in a fixed position. We used 13 Sony DXC 9000 (3 CCD, progressive scan) cameras with all gain and gamma correction turned off. We augmented the 3D Room with 21 Minolta 220X flashes controlled by an Advantech PCL-734 digital output board, duplicating the Yale "flash dome" used to capture the database in [Georghiades *et al.*, 2000].

Capturing images of every person under every possible combination of pose, illumination, and expression was not practical because of the huge amount of storage space required. The PIE database therefore consists of two major partitions, the first with pose and illumination variation, the second with pose and expression variation. There is no simultaneous variation in illumination and expression because it is more difficult to systematically vary the illumination while a person is exhibiting a dynamic expression.

In the remainder of this paper we describe the capture hardware in the CMU 3D Room, the capture procedure, the organization of the database, several possible uses of the database, and how to obtain a copy of it.

2 Capture Apparatus and Procedure

2.1 Setup of the Cameras: Pose

Obtaining images of a person from multiple poses requires either multiple cameras capturing images simultaneously, or multiple "shots" taken consecutively (or a combination of the two.) There are a number of advantages of using multiple cameras: (1) the process takes less time, (2) if the cameras are fixed in space, the (relative) pose is the same

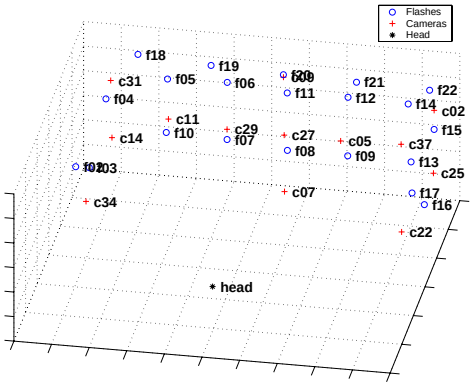


Figure 2: The xyz-locations of the head position, the 13 cameras, and the 21 flashes plotted in 3D to illustrate their relative locations. The locations were measured with a Leica theodolite. The numerical values of the locations are included in the database.

for every subject and there is less difficulty in positioning the subject to obtain a particular pose, (3) if the images are taken simultaneously we know that the imaging conditions (i.e. incident illumination, etc) are the same. This final advantage can be particularly useful for detailed geometric and photometric modeling of objects. On the other hand, the disadvantages of using multiple cameras are: (1) We actually need to possess multiple cameras, digitizers, and computers to capture the data. (2) The cameras need to be synchronized: the shutters must all open at the same time and we must know the correspondence between the frames. (3) Despite our best efforts to standardize settings, the cameras will have different intrinsic and extrinsic parameters.

Setting up a synchronized multi-camera imaging system is quite an engineering feat. Fortunately, such a system already existed at CMU, namely the 3D Room [Kanade *et al.*, 1998]. We reconfigured the 3D Room and used it to capture multiple images of each person simultaneously across pose.

Figure 1 shows the capture setup in the 3D Room. There are 49 cameras in the 3D Room, 14 very high quality (3 CCD, progressive scan) Sony DXC 9000's, and 35 lower quality (single CCD, interlaced) JVC TK-C1380U's. We decided to use only the Sony cameras so that the image quality is approximately the same across the entire database. Due to other constraints we were only able to use 13 of the 14 Sony cameras. This still allowed us to capture 13 poses of each person simultaneously however.

We positioned 9 of the 13 cameras at roughly head height in an arc from approximately a full left profile to a full right profile. Each neighboring pair of these 9 cameras are therefore approximately 22.5° apart. Of the remaining 4 cameras, 2 were placed above and below the central (frontal) camera, and 2 were placed in the corners of the room where a typical surveillance camera would be. The locations of 10 of the cameras can be seen in Figure 1. The other 3 are symmetrically opposite the 3 right-most cameras visible in the figure. Finally we measured the locations of the cameras using a theodolite. The measured locations are shown in Figure 2. The numerical values are included in the database.

The pose of a person's head can only be defined relative to a fixed direction, most naturally the frontal direction. Although this fixed direction can perhaps be defined using anatomical measurements, even this method is inevitably somewhat subjective. We therefore decided to define pose by asking the person to look directly at the center camera (c27 in our numbering scheme.) The subject therefore defines what is frontal to them. In retrospect this should have been done more precisely because some of the subjects clearly introduced an up-down tilt or a left-right twist. The absolute pose measurements that can be computed from the head position, the camera position, and the frontal direction (from the head position to camera c27) should therefore be used with caution. The relative pose, on the other hand, can be trusted. The PIE database can be used to evaluate the performance of pose estimation algorithms either by using the absolute head poses, or by using the relative poses to estimate the internal consistency of the algorithms.

2.2 The Flash System: Illumination

To obtain significant illumination variation we extended the 3D Room with a "flash system" similar to the Yale Dome used to capture the data in [Georghiades *et al.*, 2000]. With help from Athinodoros Georghiades and Peter Belhumeur, we used an Advantech PCL-734, 32 channel digital output board to control 21 Minolta 220X flashes. The Advantech board can be directly wired into the "hot-shoe" of the flashes. Generating a pulse on one of the output channels then causes the corresponding flash to go off. We placed the Advantech board in one of the 17 computers used for image capture and integrated the flash control code into the image capture routine so that the flash, the duration of which is approximately 1ms, occurs while the shutter (duration approximately 16ms) is open. We then modified the image capture code so that one flash goes off in turn for each image captured. We were then able to capture 21 images, each with different illumination, in $21/30 \approx 0.7$sec. The locations of the flashes, measured with a theodolite, are shown in Figure 2 and included in the database meta-data.

In the Yale illumination database [Georghiades *et al.*, 2000] the images are captured with the room lights switched off. The images in the database therefore do not look entirely natural. In the real world, illumination usually consists of an ambient light with perhaps one or two point sources. To obtain representative images of such cases (that are more appropriate for determining the robustness of face recognition algorithms to illumination change) we decided to capture images both with the room lights on and with them off. We decided to include the images with the room lights off to provide images for photometric stereo.

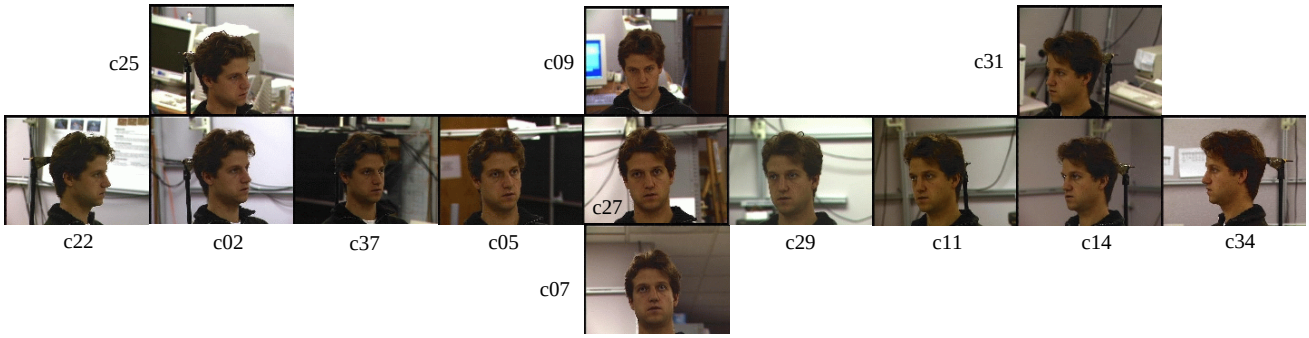


Figure 3: An illustration of the pose variation in the PIE database. The pose varies from full left profile to full frontal and on to full right profile. The 9 cameras in the horizontal sweep are each separated by about 22.5°. The 4 other cameras include 2 above and 2 below the central camera, and 2 in the corners of the room, a typical location for surveillance cameras. See Figures 1 and 2 for the camera locations.

To get images that look natural when the room lights are on, the room illumination and the flashes need to contribute approximately the same amount of light in total. The flash is much brighter, but is illuminated for a much shorter period of time. Even so, we still found it necessary to place blank pieces of paper in front of the flashes as a filter to reduce their brightness. The aperture setting is then set so that without the flash the brightest pixel registers a pixel value of around 128, while with the flash the brightest pixel is about 255. Since the "color" of the flashes is quite "hot," it is only the blue channel that ever saturates. The database therefore contains saturated data in the blue channel that is useful for evaluating the robustness of algorithms to saturation, as well as unsaturated data in both the red and green channels, which can be used for tasks that require unsaturated data, such as photometric stereo.

An extra benefit of the filtering is that the flashes are then substantially less bright than when not filtered. There are therefore no cases of the subjects either blinking or grimacing during the capture sequence, unlike in the Yale database (where the flashes are also much closer.) On the other hand, a slight disadvantage of this decision is that the images that were captured without the flashes are compressed into 0-128 intensity levels and so appear fairly dark. This can easily be corrected, but at the cost of increased pixel noise. (We found no easy way of temporally increasing the light level, or opening the aperture, for the ambient only images.)

To obtain the (pose and) illumination variation, we led each of the subjects through the following steps:

With Room Lights: We first captured the illumination variation with the room lights switched on. We asked the person to sit in the chair with a neutral expression and look at the central (frontal) camera. We then captured 24 images from each camera, 2 with no flashes, 21 with one of the flashes firing, and then a final image with no flashes. If the person wears glasses, we got them to keep them on. Although we captured this data from each camera, for reasons of storage space we decided to keep only the output of three cameras, the frontal camera, a 3/4 profile, and a full profile view.

Without Room Lights: We repeated the previous step but with the room lights off. Since these images are likely to be used for photometric stereo, we asked the person to remove their glasses if they wear them. We kept the images from all of the cameras this time. (We made the decision to keep all of the images without the room lights, but only a subset with them, to ensure that we could duplicate the results in [Georghiades *et al.*, 2000]. In retrospect we should have kept all of the images captured with the room lights on and instead discarded more images with them off.)

2.3 The Capture Procedure: Expression

Although the human face is capable of making a wide variety of complex expressions, most of the time we see faces in one of a small number of states: (1) neutral, (2) smiling, (3) blinking, or (4) talking. We decided to focus on these four simple expressions in the PIE database because extensive databases of frontal videos of more complex, but less frequently occurring, expressions are already available [Kanade *et al.*, 2000]. Another factor that effects the appearance of human faces is whether the subject is wearing glasses or not. For convenience, we include this variation in the pose and expression variation partition of the database.

To obtain the (pose and) expression variation, we led each of the subjects through the following steps:

Neutral: We asked the person to sit in the chair and look at the central camera with a neutral expression. We then captured a single frame from each camera.

Smile: We repeated the previous step, but this time asked the subject to smile.

Blink: We again repeated the previous steps, but asked the subject to close her eyes to simulate a blink.

Talking: We asked the person to look at the central camera and speak the words "1, 2, 3, . . ." while we captured 2 seconds (60 frames) of video from each camera.

Without Glasses: If the subject wears glasses, we repeated the neutral scenario, but without the glasses.

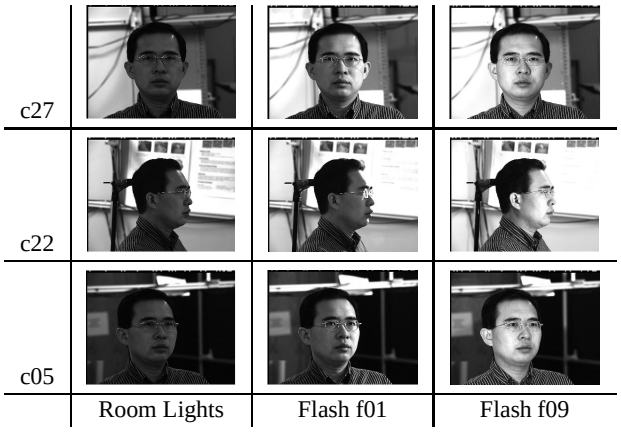


Figure 4: An example of the pose and illumination variation with the room lights on. The subject is asked to pose with a neutral expression and to look at the central camera (c27). We then capture 24 images (for each camera): 2 with just the background illumination, 21 with one of the flashes firing, and one final image with just the background illumination. Notice how the combination of the background illumination and the flashes leads to much more natural looking images than with just the flash; c.f. Figure 5.

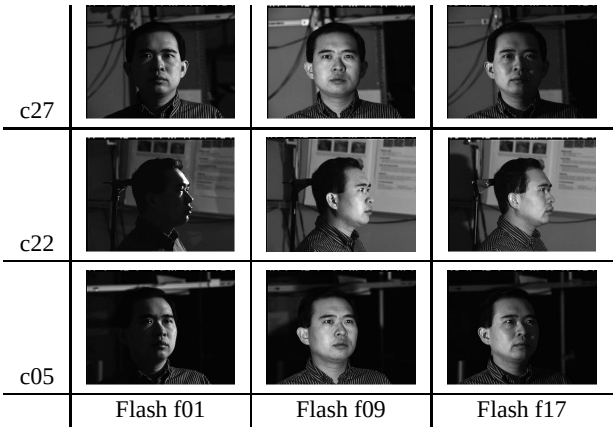


Figure 5: An example of the pose and illumination variation with the room lights off. This part of the database corresponds to the Yale illumination database [Georghiades *et al.*, 2000]. We captured it to allow direct comparison with the Yale database. This part of the database is less representative of facial images that appear in the real world than those in Figure 4 but can be used recover 3D face models using photometric stereo.

In all these steps the room lights are lit and the flash system is switched off. We also always captured images from all 13 cameras. However, because the storage requirements of keeping 60 frames of video for all cameras and all subjects is very large, we kept the "talking" sequences for only 3 cameras: the central camera, a 3/4 profile, and a full profile.

3 Database Organization

On average the capture procedure took about 10 minutes per subject. In that time, we captured (and retained) over 600 images from 13 poses, with 43 different illuminations, and with 4 expressions. The images are 640×486 color images. (The first 6 rows of the images contain synchronization information added by the VITC units in the 3D Room [Kanade *et al.*, 1998]. This information could be discarded.) The storage required per person is approximately 600MB using color "raw PPM" images. Thus, the total storage requirement for 68 people is around 40GB (which can of course be reduced by compressing the images.)

The database is organized into two partitions, the first consisting of the pose and illumination variation, the second consisting of the pose and expression variation. Since the major novelty of the PIE database is the pose variation, we first discuss the pose variation in isolation before describing the two major partitions. Finally, we include a description of the database meta-data (i.e. calibration data, etc.)

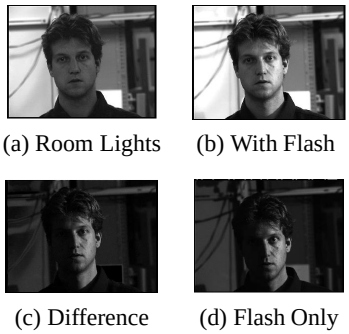
3.1 Pose Variation

An example of the pose variation in the PIE database is shown in Figure 3. This figure contains images of one subject in the database from each of the 13 cameras. As can be seen, there is a wide variation in pose from full profile to full frontal. This subset of the data should be useful for evaluating the robustness of face recognition algorithms across pose. Since the camera locations are known, it can also be used for the evaluation of pose estimation algorithms. Finally, it might be useful for the evaluation of algorithms that combine information from multiple widely separated views. An example of such an algorithm would be one that combines frontal and profile views for face recognition.

3.2 Pose and Illumination Variation

Examples of the pose and illumination variation are shown in Figures 4 and 5. Figure 4 contains the variation with the room lights on and Figure 5 with the lights off. Comparing the images we see that those in Figure 4 appear more natural and representative of images that occur in the real world. On the other hand, the data with the lights off was captured to reproduce the Yale database [Georghiades *et al.*, 2000]. This will allow a direct comparison between the two databases. Besides the room lights, the other major differences between these parts of the database are: (1) the subjects wear their glasses in Figure 4 (if they have them) and not in Figure 5, and (2) in Figure 5 we retain all of the images, whereas for Figure 4 we only keep the data from 3 cameras, the frontal camera c27, the 3/4 profile camera c22, and the full profile camera c05. We foresee a number of possible uses for the pose and illumination variation data. First it can be used to reproduce the results in [Georghiades *et al.*, 2000]. Secondly it can be used to evaluate the robustness of face recognition algorithms to pose and illumination.

A natural question that arises is whether the data with the



(a) Room Lights (b) With Flash

(c) Difference (d) Flash Only

Figure 6: An example of an image with room lights and a single flash (b), and subtracting from it an image with only the room lights (a) taken a fraction of a second earlier. The difference image (c) is compared with an image taken with the same flash but without room lights (d). Although the facial expression is a little different, the images otherwise appear similar. (There are also a number of differences in the background caused by certain pixels saturating when the flash is illuminated.)

room lights on can be converted into that without the lights by simply subtracting an image with no flash (but with just the background illumination) from images with both. Preliminary results indicate that this is the case. For example, Figure 6 contains an image with just the room lights and another image taken with both the room lights and one of the flashes a short fraction of a second later. We show the difference between these two images and compare it with an image of the same person taken with just the flash; i.e. with the room lights off. Except for the fact that the person has a slightly different expression (that image was taken a few minutes later), the images otherwise look fairly similar. We have yet to try to see whether vision algorithms behave similarly on these two images. If they do, we can perhaps form synthetic images of a person captured under multiple flashes and add them to the database.

3.3 Pose and Expression Variation

An example of the pose and expression variation is shown in Figure 7. The subject is asked to provide a neutral expression, to smile, to blink (i.e. they are asked to keep their eyes shut), and to talk. For neutral, smiling, and blinking, we kept all 13 images, one from each camera. For talking, we captured 2 seconds of video (60 frames.) Since this occupies a lot more space, we kept this data for only 3 cameras: the frontal camera c27, the 3/4 profile camera c22, and the full profile camera c05. In addition, for subjects who usually wear glasses, we collected one extra set of 13 images without their glasses (and with a neutral expression.)

The pose and expression variation data can possibly be used to test the robustness of face recognition algorithms to expression (and pose.) A special reason for including blinking was because many face recognition algorithms use the eye pupils to align a face model. It is therefore possible

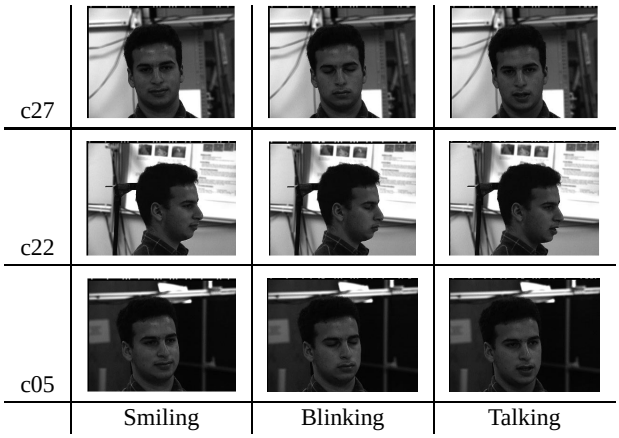


Figure 7: An example of the pose and expression variation in the PIE database. Each subject is asked to give a neutral expression (image not shown), to smile, to blink, and to talk. We capture this variation in expression across all poses. For the neutral images, the smiling images, and the blinking images, we keep the data for all 13 cameras. For the talking images, we keep 60 frames of video from only three cameras (frontal c27, 3/4 profile c05, and full profile c22). For subjects who wear glasses we also capture one set of 13 neutral images of them without their glasses.

that they are particularly sensitive to subjects blinking. We can now test whether this is indeed the case.

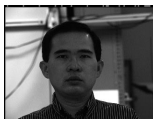
3.4 Meta-Data

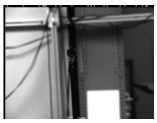
Besides the two major partitions of the database, we also collected a variety of miscellaneous "meta-data" to aid in calibration and other processing:

Head, Camera, and Flash Locations: Using a theodolite, we measured the xyz-locations of the head, the 13 cameras, and the 21 flashes. See Figure 2 for an illustration. The numerical values of the locations are included in the database and can be used to estimate (relative) head poses and illumination directions.

Background Images: At the start of each recording session, we captured a background image from each of the 13 cameras. An example is shown in Figure 8(b). These images can be used for background subtraction to help localize the face region. As can be seen in Figure 8(c), background subtraction works very well. The head region can easily be segmented in Figure 8(c). Because the subject doesn't move, background subtraction can also be performed between the neutral image and the background image to create a mask that can be used with all the illumination variation images captured with the room lights on. (See Figure 4.) No background images are provided for the images captured with the room lights off. (See Figure 5.)

Color Calibration Images: Although the cameras that we used are all of the same type, there is still a large


(a) PIE Image


(b) Background Image


(c) Difference

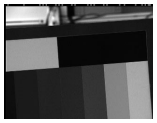

(d) Color Calibration Image

Figure 8: An example of a background image (b) and a demonstration of how background subtraction can be used to locate the face (c). This may be useful in evaluations where we do not want to evaluate localization. An example color calibration image (d). These images can be used to estimate simple linear response functions for each of the color channels to color calibrate the cameras.

amount of variation in their photometric responses, both due to their manufacture and due to the fact that the aperture settings on the cameras were all set manually. We did "auto white-balance" the cameras, but there is still some noticeable variation in their color response. To allow the cameras to be intensity- (gain and bias) and color-calibrated, we captured images of color calibration charts at the start of every session and include them in the database meta-data. Although we do not know "ground-truth" for the colors, the images can be used to equalize the color (and intensity) responses across the 13 cameras. An example of a color calibration image is shown in Figure 8(d).

Personal Attributes of the Subjects: Finally, we include some personal information about the 68 subjects in the database meta-data. For each subject we record the subject's sex and age, the presence or absence of eye glasses, mustache, and beard, as well as the date on which the images were captured.

4 Potential Uses of the Database

Throughout this paper we have pointed out potential uses of the database. We now summarize some of the possibilities:

- Evaluation of head pose estimation algorithms.
- Evaluation of the robustness of face recognition algorithms to the pose of the probe image.
- Evaluation of face recognition algorithms that operate across pose; i.e. algorithms for which the gallery and probe images have different poses.
- Evaluation of face recognition algorithms that use multiple images across pose (gallery, probe, or both).
- Evaluation of the robustness of face recognition algorithms to illumination (and pose).
- Evaluation of the robustness of face recognition algorithms to common expressions (and pose).
- 3D face model building either using multiple images across pose (stereo) or multiple images across illumination (photometric stereo [Georghiades *et al.*, 2000]).

Although the main uses of the PIE database are for the evaluation of algorithms, the importance of such evaluations (and the databases used) for the development of algorithms should not be underestimated. It is often the failure of existing algorithms on new datasets, or simply the existence of new datasets, that drives research forward.

5 Obtaining the Database

Because the PIE database (uncompressed) is over 40GB, we have been distributing it in the following manner:

1. The recipient ships an empty (E)IDE hard drive to us.
2. We copy the data onto the drive and ship it back.

To date we have shipped the PIE database to over 20 research groups worldwide. Anyone interested in receiving the database should contact the second author by email at simonb@cs.cmu.edu or visit the PIE database web site at http://www.ri.cmu.edu/projects/project_418.html.

Acknowledgements

We would like to thank Athinodoros Georghiades and Peter Belhumeur for giving us the details the Yale "flash dome." Sundar Vedula and German Cheung gave us great help using the CMU 3D Room. We would also like to thank Henry Schneiderman and Jeff Cohn for discussions on what data to collect and retain. Financial support for the collection of the PIE database was provided by the U.S. Office of Naval Research (ONR) under contract N00014-00-1-0915. Finally, we thank the FG 2002 reviewers for their feedback.

References

[Georghiades *et al.*, 2000] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Generative models for recognition under variable pose and illumination. In *Proc. of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[Kanade *et al.*, 1998] T. Kanade, H. Saito, and S. Vedula. The 3D room: Digitizing time-varying 3D events by synchronized multiple video streams. Technical Report CMU-RI-TR-98-34, CMU Robotics Institute, 1998.

[Kanade *et al.*, 2000] T. Kanade, J. Cohn, and Y.-L. Tian. Comprehensive database for facial expression analysis. In *Proc. of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[Philips *et al.*, 1997] P.J. Philips, H. Moon, P. Rauss, and S.A. Rizvi. The FERET evaluation methodology for face-recognition algorithms. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 1997.